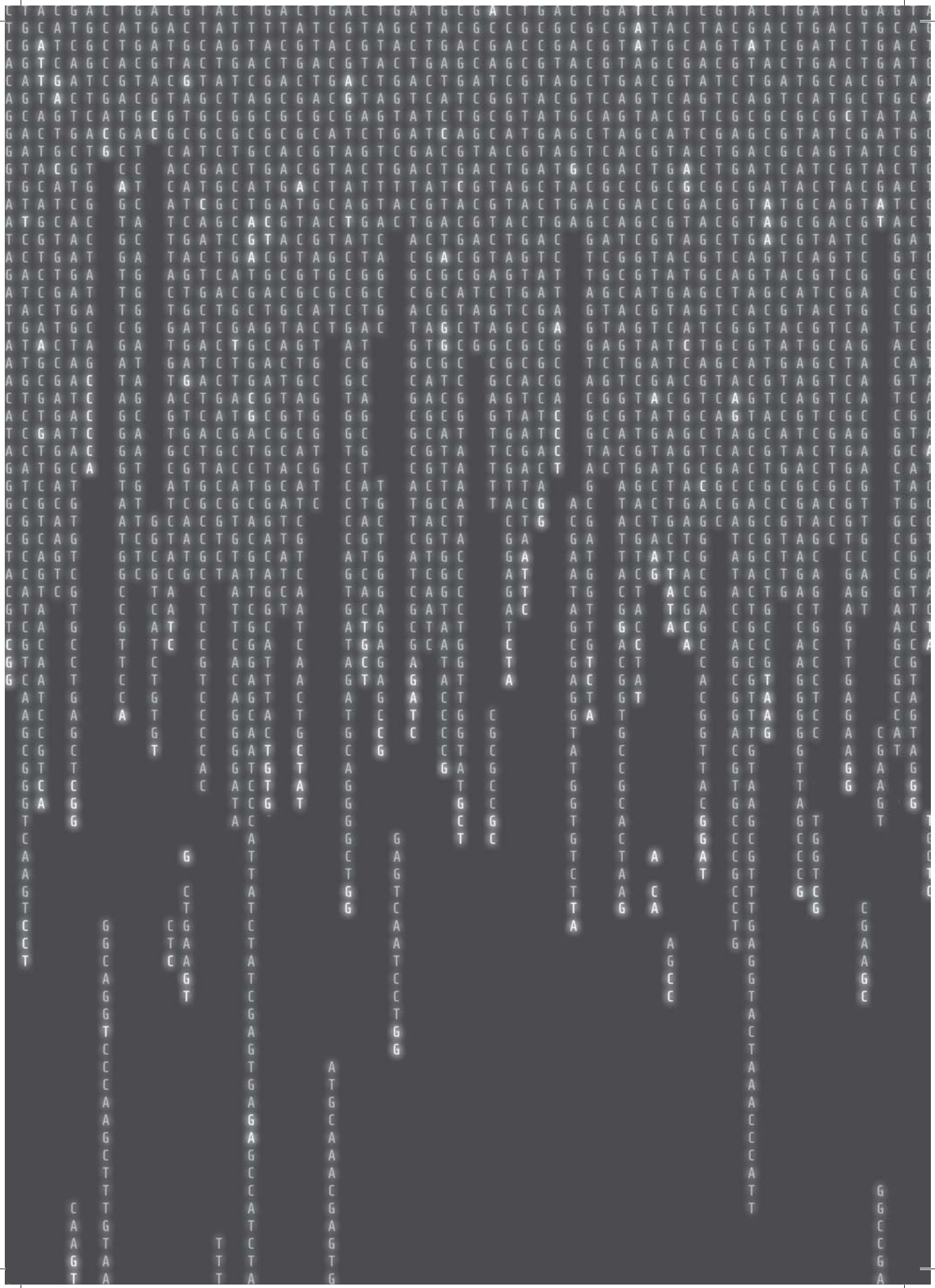
Cracking the code-ing sequence for Parkinson’s disease

Jansen, I.E.

2017

document version
Publisher's PDF, also known as Version of record



CHAPTER 2

CHCHD2: A RARE GENETIC RISK FACTOR OF PARKINSON'S DISEASE IN THE CAUCASIAN POPULATION

A modified version of this chapter has been published as: Jansen IE, Bras JM, Lesage S, Schulte C, Gibbs JR, Nalls MA, et al. CHCHD2 and Parkinson's disease. *Lancet Neurology*, 2015; 14(7): 678-9.

ABSTRACT

The gene *CHCHD2* has recently been identified as a novel genetic factor causing Parkinson's disease (PD) in the Asian population. To determine the influence of *CHCHD2* variants in the Western European population, we explored two genetic exome datasets of the International Parkinson's Disease Genomics Consortium (IPDGC), including a whole exome sequencing dataset (WES: 1,243 PD cases and 472 controls) and an array-based exome dataset (NeuroX: 6,927 PD cases and 6,108 controls). We were unable to replicate the rare variants that were detected in the Asian population, suggesting that they are very rare and Asian-specific rather than a common genetic basis for Parkinson's disease. The identification of 3 distinct coding rare variants that are predicted to be pathogenic in 4 PD cases of our WES dataset implies that *CHCHD2* is also a rare genetic risk factor in Western European population. Future large-scale sequencing studies will further establish the genuine influence of *CHCHD2* on PD pathogenesis.

INTRODUCTION

Genetic discoveries have lent support to the growing consensus that protein aggregation, impairment in oxidative stress, and mitochondrial dysfunction are important pathways in Parkinson's disease (PD).[1] In agreement with this hypothesis, Funayama and colleagues[2] report a heterozygous mutation (p.Thr61Ile) in *CHCHD2*, which is suggested to be involved in mitochondrial respiration. This variant was detected in a large multi-generation Japanese family with an autosomal dominant form of PD. The screening of additional 330 index cases with autosomal dominant PD observed another nonsynonymous (p.Arg145Gln) and a splice site variant (c.300+5G>A). Likewise, sequencing of sporadic PD cases identified a number of putative risk variants in *CHCHD2* in the same population.

Novel genetic factors such as *CHCHD2* require replication, both within the discovery population and in different ethnic populations. Genotyping the observed Asian variants would be a cost-effective option, yet other mutations are missed. As the identified variants are rare and observed in a non-European population, it is important to explore for other likely pathogenic mutations, in addition to the identical variants. Therefore, sequencing data is preferred. Whole exome sequencing (WES) data is a great source to further investigate novel genes recently published in the genomics research field. The WES dataset of the International Parkinson's Disease Genomics Consortium is of sufficient size to perform valuable examinations to determine the extent of replication. We used this dataset, together with data generated by the array-based exome NeuroX chip, to establish the pathogenicity of *CHCHD2* in individuals of western European ancestry.

METHODS

Whole exome sequencing dataset

We investigated the presence and genetic burden of putative pathogenic rare variants in *CHCHD2* by exploring the IPDGC WES dataset comprising 1,243 PD cases (average age of onset 42.3 years) and 472 controls. Sample libraries were prepared with Roche Nimblegen or Illumina exome capture kits, and subjected to 100-bp paired-end sequencing on the Illumina HiSeq2000. The reads were aligned using BWA-MEM 0.7.9a to the reference genome (UCSC hg19).[3] GATK 3.x and ANNOVAR were used to call, quality-based filter, and annotate variants.[4,5] For functional predictions, 6 different algorithms were used: SIFT[6], PhyloP[7], Polyphen[8], LRT[9], MutationTaster[10] and CADD[11]. Variants were only considered for subsequent analyses if targeted by both capture kits.

NeuroX exome array dataset

The dataset generated by the exome genotyping array NeuroX, consisting of 6,927 cases and 6,108 controls, was used to investigate the role of putative risk variants in *CHCHD2*.

Table 1. *CHCHD2* variants in IPDGC cohort.

	Change		dbSNP141	ExAC		Path.*	Funayama paper◦	
	cDNA	AA		EUR	EAS	pred.	Fam. PD	Spor. PD
Position (hg19)	(NM_016139)						(n=340)	(n=517)
chr7:56174129	-23G>A	UTR5	rs368503740	0·00006	0·0	NA	··	··
chr7:56174118	-12C>T	UTR5	rs112876794	0·0008	0·0	NA	··	··
chr7:56174102	5C>T	Pro2Leu	rs142444896	0·004	0·008	b,d,e,f	0·035	0·018
chr7:56174067	40C>T	Pro14Ser	rs137965562	0·0001	0·005	b,e,f	··	··
chr7:56172125	94G>A	Ala32Thr	rs145190179	0·00006	0·0006	b,d,e,f	··	··
chr7:56172118	101C>T	Pro34Leu	rs371198317	0·0003	0·0	a,b,d,e,f	··	··
chr7:56171981	238A>G	Ile80Val	rs149119842	0·00003	0·0	f	··	··
chr7:56169498	+46Tins	UTR3	rs35957514	0·0002	0·0	NA	··	··
chr7:56169469	+75T>C	UTR3	unknown	ND	ND	NA	··	··

ExAC=Exome Aggregation Consortium; AA=amino acid; EUR=European population; EAS=East Asian population; Fam.=familial; PD=Parkinson's disease; Spor.=sporadic; Contr.=controls; OR=odds ratio; UTR=untranslated region; ins=insertion; ND=no data available in database; NA=not applicable for non-exonic variants; ··=variant not present in this dataset. *Path. pred. displays which of algorithms predict pathogenicity: a=SIFT[6], b=PhyloP[7], c=Polyphen2[8], d=LRT[9], e=MutationTaster[10], f=CADD[11]. ◦Only one exact same variant in Funayama paper and IPDGC datasets. ▪Logistic regression adjusts for first 4 multi-imensional scaling components and gender.

The NeuroX array[12] contains approximately 240,000 variants of standard Illumina exome content and approximately 24,000 variants implied to be associated to neurodegenerative diseases.

Genetic analyses

Both for the WES and the NeuroX dataset, variants were only considered when passing standard variant quality control steps (variant exclusion when missing genotype rate > 5%, Hardy Weinberg equilibrium p-value < $1e^{-6}$ and non-random missingness by phenotype p-value < $1e^{-5}$). Minor allele frequencies were determined per individual variant. Statistical comparisons were performed if applicable, which depends on the sample size and variant frequency. For the exome dataset, the sequence kernel association test (SKAT)[13] was applied to perform a gene-based association analysis. This rare variant aggregation association test accounted for the putative confounding factors gender, country of origin, capture kit type, portion of the exome that has been captured for at least 10x and the first 4 MDS components. The SKAT analysis was performed on all variants identified in *CHCHD2* or coding variants only. For the NeuroX dataset, logistic regression was performed that determined the effect estimate of two *CHCHD2* coding variant, while correcting for covariates gender, country of origin and the first 4 MDS components.

Funayama paper◦			IPDGC					
Contr.	Fisher exact		Exome sequencing		NeuroX			
(n=559)	OR	*p*	PD (n=1243)	Contr. (n=472)	PD (n=6927)	Contr.(n=6108)	OR	*p*▪
..	..	..	0·0004	0	..	..	..	..
..	..	..	0·0008	0	..	..	..	..
0·004	4·69	0·0025	..	..	0·0004	0·0003	1·096	0·889
..	..	..	..	..	0·0004	0·0003	1·281	0·736
..	..	..	0·0004	0	..	..	..	..
..	..	..	0·0008	0	..	..	..	..
..	..	..	0·0004	0	..	..	..	..
..	..	..	0	0·001	..	..	..	..
..	..	..	0·001	0	..	..	..	..

Linkage-disequilibrium calculations

The VCF files of the Thousand Genomes Project Phase 3[14] covering *CHCHCD2* were downloaded per population through the online ensemble browser.[15] The European population and East Asian populations comprised 503 and 504 unrelated individuals, respectively. VCFtools[16] v0.1.11 converted the VCF files into PLINK format to make them compatible with Haploview[17] 4.2. The standard LD color scheme (D'/LOD) was used in Haploview to visualize the LD plot. For the completeness of the picture, all variants were considered, even when the minor allele frequency was 0.

RESULTS

Table 1 shows *CHCHD2* variants identified in our exome and NeuroX datasets. WES captured 8 variants that were located in regions targeted by both capture kits, including three coding variants, three 5' UTR variants and two 3' UTR variants. One variant in the 5' UTR region was removed due to significant difference in missingness between cases and controls. We did not note any of the putative causal variants (p.Thr61Ile, p.Arg145Gln, and c.300+5G>A) described by Funayama et al.[2] Also, these variants are absent in the European population in the publicly accessible Exome Aggregation Consortium (ExAC) browser,[18] suggesting that they are very rare and Asian-specific rather than a common genetic basis for Parkinson's disease. However, in our dataset, among people of western European ancestry, three novel putative pathogenic variants in exon 2 (p.Ala32Thr, p.Pro34Leu, and p.Ile80Val) were noted in four of 1,243 cases (<1%). These variants are predicted to be

pathogenic (Table 1) and the affected amino acids are conserved down to rodents. This conserved sequence together with the lack of identification of exonic variants in controls indicate that variation in *CHCHD2* might be a rare risk factor in people of western European ancestry. The burden analysis of rare variants did not show association with Parkinson's disease in the exome dataset, neither considering the 3 exonic variants only (p = 0.28), nor when testing for the joint effect of all 7 variants including the UTR variants (p = 0.24).

The four patients with Parkinson's disease with three novel variants were two individuals from the USA and two from France. The US cases were diagnosed at age 39 years and 51 years with asymmetric onset and showed typical symptoms such as bradykinesia and, in one patient, resting tremor. Although these two patients were reported to have familial Parkinson's disease, we did not have access to family members to find out whether the variant cosegregated with the disease. The two French patients were isolated cases with an age at onset of 20 years and 39 years.

Using NeuroX data, we captured two variants in our Caucasian population (Table 1). One of these variants, rs142444896, was reported by Funayama and colleagues[2] to be significantly associated with Parkinson's disease in the Japanese population. Although we estimated that our NeuroX design has 99% power to detect variants with an allele frequency of 0.4% (European frequency in ExAC) and an odds ratio of 4.69, we could not replicate (p = 0.89) this association within our western European population. The second variant, not observed in the original *CHCHD2* publication, also lacked an association to PD (p = 0.74).

We investigated the linkage-disequilibrium (LD) patterns of the two different populations to determine whether the variants observed in our study are independent loci from the Japanese ones, especially for the more common NeuroX variant that was not reported by Funayama et al.[2] Figure 1 shows the LD blocks for all variants identified by Funayama et al. and this study, which were present within the publically available 1000 Genomes Project dataset. The patterns are similar for the different populations. As expected, none of the rare variants are in LD. Only 3 variants (rs8406, rs1244896 and rs200226056) seem to be inherited together, which are the 3 mutations that showed significant association to sporadic PD in the Japanese population. The one NeuroX variant that was only reported by our study, rs137965562, is not in strong LD with any variant observed by Funayama et al, neither in the 1000 G European nor the East-Asian population.

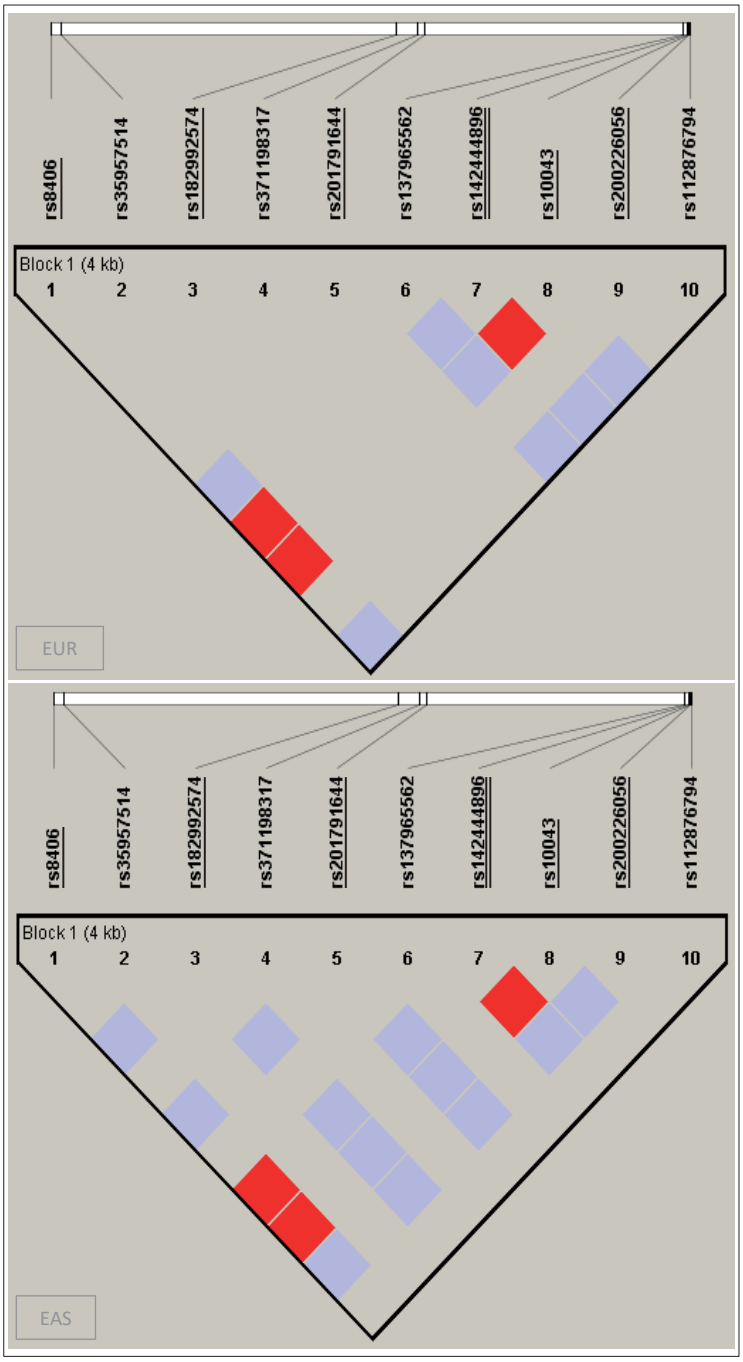


Figure 1. LD patterns of European (EUR) and East-Asian (EAS) 1000G data. Red blocks indicate a D'-score of 1 and a LOD ≥ 2. The SNPs underlined were reported by Funayama et al. The double underlined SNP was present in both Funayama et al and our study.

DISCUSSION

By exploring two independent genetic exome datasets of the IPDGC, we aimed to establish the influence of *CHCHD2* variants in Parkinson's disease patients with western European ancestry. We identify three nonsynonmous variants that are predicted to have a damaging effect on the function of the gene. These putative risk variants in *CHCHD2* emphasize the current thinking that mitochondrial dysfunction plays an important part in PD. In this regard, dysregulation of PINK1/parkin pathway, which is especially associated to young-onset PD, might result in damaged mitochondria.[1] Interestingly, three patients carrying an exonic *CHCHD2* variant in our cohort were diagnosed with PD at an early age (< 45 years).

The frequency of newly identified putative variants in our exome sequencing dataset ranged from 0.04% to 0.08%, higher than that reported in ExAC (0.003% to 0.03%),[18] which suggest that they are likely to be rare risk variants for PD in populations of European ancestry. However, there was no association for the two more common variants that were captured by the NeuroX genotyping array. One of these two NeuroX variants, rs137965562, was reported as significantly associated to PD by Funayama et al. Our LD pattern analyses revealed that this variant is on the same haplotype as variants rs8406 and rs1244896, both in the European as the East-Asian population. It is therefore highly probable, that these two additional variants would also lack an association in our IPDGC NeuroX cohort, if they would have been genotyped. We therefore conclude that in our IPDGC cohort *CHCHD2* is not a common risk factor for PD, which is in agreement with the results of a Japanese GWAS and the most recent and largest PD meta-analysis.[19,20]

The identification of 3 rare possibly pathogenic mutations in our PD cohort suggests that *CHCHD2* in the Caucasian population is a rare cause for PD, including young-onset PD. However, the 3 *CHCHD2* variants in the familial cases of Funayama et al., seem to be very rare genetic causes of PD and are most likely specific for the East Asian population. Because the association of common *CHCHD2* variants of sporadic cases was unable to replicate within our European NeuroX dataset, it suggests that *CHCHD2* is not a common risk factor for PD. To further provide evidence to substantiate the role of *CHCHD2* in PD, more sequencing studies of *CHCHCD2* in multiple ethnic populations should be encouraged. At least for the European population it occurs that the more rare variants have potential to influence PD pathogenesis, meaning large genetic datasets of ten thousands of individuals are anticipated to be needed for adequate replication.

REFERENCES

1. Trinh J, Farrer M. Advances in the genetics of Parkinson disease. *Nature reviews Neurology* 2013; 9(8): 445-54.
2. Funayama M, Ohe K, Amo T, et al. CHCHD2 mutations in autosomal dominant late-onset Parkinson's disease: a genome-wide linkage and sequencing study. *Lancet neurology* 2015.
3. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* 2009; 25(14): 1754-60.
4. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* 2010; 20(9): 1297-303.
5. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* 2010; 38(16): e164.
6. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* 2009; 4(7): 1073-81.
7. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research* 2010; 20(1): 110-21.
8. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nature methods* 2010; 7(4): 248-9.
9. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome research* 2009; 19(9): 1553-61.
10. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature methods* 2010; 7(8): 575-6.
11. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM. A general framework for estimating the relative pathogenicity of human genetic variants. 2014; 46(3): 310-5.
12. Nalls MA, Bras J, Hernandez DG, et al. NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases. *Neurobiology of aging* 2014.
13. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *American journal of human genetics* 2011; 89(1): 82-93.
14. Auton A, Brooks LD, Durbin RM, et al. A global reference for human genetic variation. *Nature* 2015; 526(7571): 68-74.
15. Yates A, Akanni W, Amode MR, et al. Ensembl 2016. *Nucleic acids research* 2016; 44(D1): D710-6.
16. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics (Oxford, England)* 2011; 27(15): 2156-8.
17. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics (Oxford, England)* 2005; 21(2): 263-5.
18. Exome Aggregation Consortium (ExAC) C, MA (URL: http://exac.broadinstitute.org) [April 2015]. 2015.
19. Satake W, Nakabayashi Y, Mizuta I, et al. Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nature genetics* 2009; 41(12): 1303-7.
20. Nalls MA, Pankratz N, Lill CM, et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nature genetics* 2014; 46(9): 989-93.